グラフィカルラッソの
糖尿病診断への適用性に関する検証

金　ヨンキョン[1]・尹　禮分[2]・Min Yoon[3]・中山　弘隆[4]

[1] 学生会員　関西大学　理工学研究科都市システム工学分野（〒564-8680 大阪府吹田市山手町3-3-35）
E-mail:k545649@kansai-u.ac.jp

[2] 正会員　関西大学教授　環境都市工学部都市システム工学科（〒564-8680 大阪府吹田市山手町3-3-35）
E-mail:yeboon@kansai-u.ac.jp

[3] 非会員　Professor of Pukyong National University（45 Yongso-ro, Nam-gu, Busan 48513, Korea）
E-mail:myoon@pknu.ac.kr

[4] 非会員　Emeritus Professor of Konan University
E-mail:nakayama@konan-u.ac.jp

本研究では，疎構造学習に基づく異常検知または変化検知を抽出する手法の一つであるグラフィカルラッソを用いて異なる2つのデータ集合間の特徴を分析する．Pima Indians Diabetesデータを対象とし，健常者と糖尿患者のデータに対する隣接行列からなる相関グラフを比較することで，要因間の相関変化を把握する．また，その結果に基づく要因の相関変化度を定量的に評価し，データの構造変化に影響のある重要要因の検出を試みる．また，サポートベクターマシンを用いて各要因の寄与度を検証し，健常者と糖尿患者の識別にあたっての重要要因を調べ，相関変化度に基づく重要要因との違いを確認する．

Key Words : *graphical lasso, change detection, support vector machine*

1. はじめに

不規則的な食習慣と睡眠パターンが主な原因である生活習慣病の代表的なものは高血圧，糖尿病，認知症がある．2015年日本人口動態統計[1]によると，死亡者数の約6割が生活習慣病に起因している．一方，日本ではデータヘルス計画[2]に取り組み，様々なヘルスデータを数理科学的方法により分析することで，健康増進，予防モデルの構築，病気の早期発見などにつなげる．

そこで，本研究ではPima Indians Diabetes[3]（本稿では，単に糖尿病という）データを対象とし，疎構造学習に基づく異常検知または変化検知の一手法であるグラフィカルラッソ（Graphical Lasso）[4]を用いて分析を行う．まず，異なる決定変数をもつ2つのデータ集合，「健常者」と「糖尿病者」データに対するそれぞれの隣接行列から，各要因（変数）間の相関変化を把握することで，データ集合間の特徴を分析する．また，要因の相関変化度に基づき，構造変化に影響のある重要要因を検出する．さらに，サポートベクターマシンを用いて糖尿病診断により重要な要因を調べる．その結果，病気の予防や健康増進における効率化を図り糖尿病にかかわる重要と思われる要因の抽出が可能かの検討を行う．

2. グラフィカルラッソ（graphical lasso）

データ集合として，m 次元の ℓ 個の観測値からなる

$$D=\left\{\boldsymbol{x}^1,\boldsymbol{x}^2,\cdots,\boldsymbol{x}^\ell\right\}$$

を考える．ここで，データの各次元ごとに標準化変換を行うことで一般性を失わず，精度行列 Λ を用いたm 次元の多変量正規分布は以下のように表すことができる：

$$\mathrm{N}(\boldsymbol{x}\mid 0,\Lambda^{-1})=\frac{\det\Lambda^{1/2}}{(2\pi)^{m/2}}\exp\left(-\frac{1}{2}\boldsymbol{x}^{\boldsymbol{T}}\Lambda\boldsymbol{x}\right)\qquad(1)$$

上式 (1) における精度行列 Λ は，次の事後確率を最大化することにより推定される：

$$\Lambda^{*}=\arg\max_{\Lambda}\ (\log\det\Lambda-\mathrm{tr}(S\Lambda)-\rho||\Lambda||_1)\qquad(2)$$

ここで，S はデータの標本の分散共分散行列であり，$\rho>0$ は正則化パラメータである．

グラフィカルラッソ[4]は，精度行列 Λ を隣接行列とみなし，L_1正則化項 $||\Lambda||_1$ の導入により，疎な隣接行列を

推定し，その結果変数間のスパース（Sparse）な依存関係が推定できる．

例えば，隣接行列 $\Lambda = (\lambda_{ij})$ に対し，$\lambda_{ij} = 0$ であれば変数 x_i と x_j は独立であり，隣接行列の要素が非ゼロ $\lambda_{ij} \neq 0$ となる変数間は（直接）相関を持ち，グラフ上で線を引くことで，互いの依存関係をみることができる（**図-1**）．

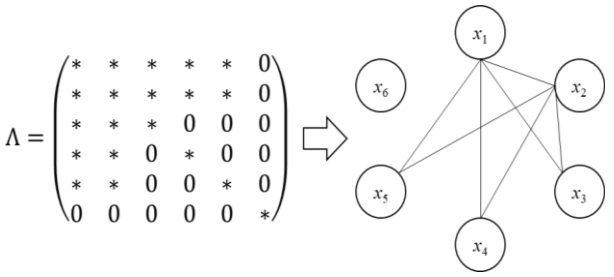
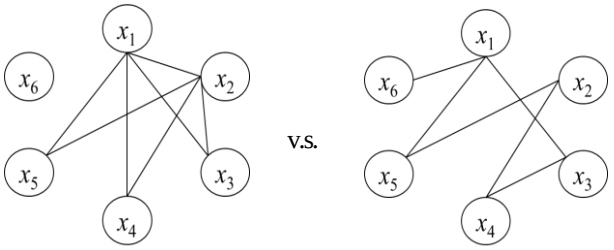


図-1　隣接行列に基づく変数間の相関図

そこで，本研究ではグラフィカルラッソによる疎な相関グラフの学習法を用いて，異なる2つのデータ集合間の特徴を分析する．それぞれのデータ集合に対する隣接行列から，各要素（変数）間の相関変化を把握することで，データ集合間の特徴を分析する（**図-2**）．さらに，変数の相関変化度[4]を算出し，構造変化に影響のある重要要因を検出する．



(a) 健常者　　(b) 患者

図-2　疎構造学習のグラフィカルラッソに基づく構造変化

3. 分析結果および考察

本研究では，Pima Indians Diabetesデータ[3] を用いて検証する．このデータは，21歳以上の女性を対象として

- Pregnancies (x_1)
- Glucose (x_2)
- Blood Pressure (x_3),
- Skin Thickness (x_4)
- Insulin (x_5)
- BMI (x_6)
- Diabetes Pedigree Function (x_7)
- Age (x_8)

の8要因に対する検査値であり，「健常者」に関するデータ $\mathrm{D_A}$ の数は 500 個であり，一方「糖尿病者」に関するデータ $\mathrm{D_B}$ の数は 268 個である．

まず，「健常者」データ $\mathrm{D_A}$ と「糖尿病者」データ $\mathrm{D_B}$ における各要因の平均差をみるため，t－検定を行い，その結果を**表-1**に示す．t－検定による結果から，Blood Pressure 以外の要因においては，「健常者」データ $\mathrm{D_A}$ と「糖尿病者」データ $\mathrm{D_B}$ の平均値に違いがあることがわかる．

表-1　t－検定の結果

要因		t 値	p 値
Pregnancies	x_1	-5.91	0.00
Glucose	x_2	-13.75	0.00
Blood Pressure	x_3	-1.71	0.04
Skin Thickness	x_4	-1.97	0.02
Insulin	x_5	-3.30	0.00
BMI	x_6	-8.62	0.00
Diabetes Pedigree Function	x_7	-4.58	0.00
Age	x_8	-6.92	0.00

次に，「健常者」データ $\mathrm{D_A}$ と「糖尿病者」データ $\mathrm{D_B}$ に対する相関分析を行い，要因間の関係性をみる．**表-2**で対角成分より上半分が「健常者」データに対する相関係数を表し，下半分が「糖尿病者」データの相関係数である．Pregnancies (x_1) と Age (x_8)，Skin Thickness (x_4) とInsulin (x_5)については、「健常者」データと「糖尿病者」データの両方の相関係数が 0.4 以上であり，これらの項目間には相関があると考えられる．一方，Skin Thickness (x_4) とBMI (x_6)に対しては，「健常者」データではある程度の相関がみられるが，これに比べて「糖尿病者」データでは相関係数が小さくなっている．

表-2　要因間の相関係数

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1		0.10	0.13	-0.12	-0.13	0.02	-0.08	0.57
x_2	-0.05		0.19	0.02	0.35	0.13	0.10	0.23
x_3	0.13	0.07		0.19	0.07	0.36	0.03	0.21
x_4	-0.08	0.04	0.23		0.41	0.44	0.10	-0.16
x_5	-0.08	0.26	0.09	0.46		0.25	0.23	-0.15
x_6	-0.16	0.05	0.13	0.31	0.06		0.07	0.04
x_7	-0.07	0.03	0.03	0.27	0.10	0.14		0.04
x_8	0.44	0.10	0.26	-0.09	0.02	-0.19	-0.09	

式 (2) における正則化パラメータを $\rho = 0.2$ とし，「健常者」データ $\mathrm{D_A}$ および「糖尿病者」データ $\mathrm{D_B}$ に対する精度行列 Λ_A および Λ_B を求め，その結果に基づく相関図を**図-3**と**図-4**に示す．ここで，「健常者」データ $\mathrm{D_A}$ に対して相関はあるが，「糖尿病者」データ $\mathrm{D_B}$ においては無（直接）相関になる要因間を，**図-3**で点線で表す．**図-4**の太線は，その逆の関係を示している．

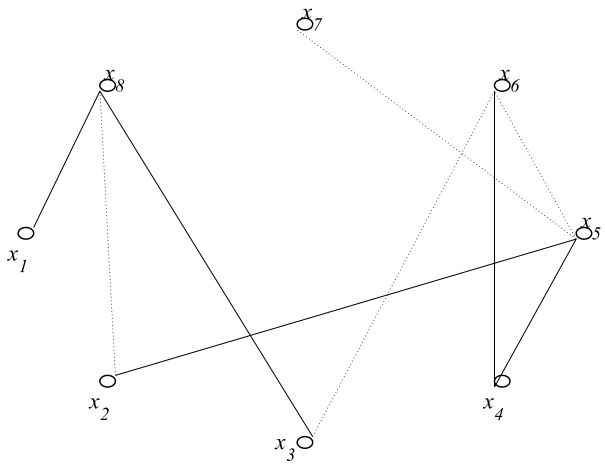


図-3 「健常者」データ $\mathrm{D_A}$ に対する相関グラフ

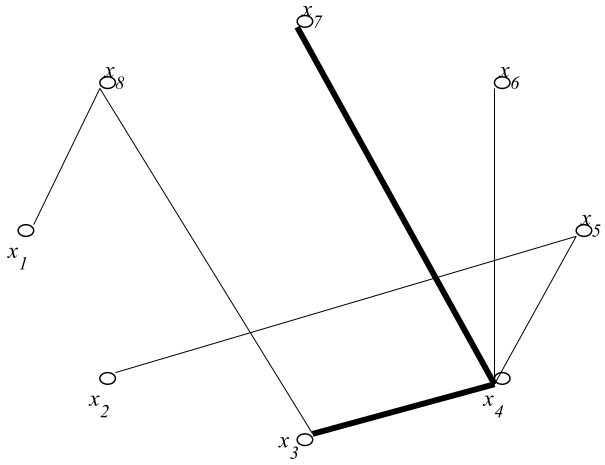


図-4 「糖尿病者」データ $\mathrm{D_B}$ に対する相関グラフ

さらに，「健常者」データ $\mathrm{D_A}$ と「糖尿病者」データ $\mathrm{D_B}$ の間の変化をはかるために，要素間の変化を期待 Kullback-Leibler（KL）距離 [4] による相違度を測り，要因ごとの変化を定量的に評価する：

$$a_i := \frac{1}{2}\ln\frac{[\Lambda_A]_{ii}}{[\Lambda_B]_{ii}} - \frac{1}{2}\left\{\frac{[\Lambda_A S \Lambda_A]_{ii}}{[\Lambda_A]_{ii}} - \frac{[\Lambda_B S \Lambda_B]_{ii}}{[\Lambda_B]_{ii}}\right\} \qquad (3)$$

上式 (3) から算出した「健常者」データと「糖尿病者」データにおける各要因の相関変化度を**表-3**に示す．

表-3 要因ごとの相関変化度

要因		変化度 a_i
Pregnancies	x_1	0.04
Glucose	x_2	0.02
Blood Pressure	x_3	0.03
Skin Thickness	x_4	0.01
Insulin	x_5	0.01
BMI	x_6	0.06
Diabetes Pedigree Function	x_7	0.01
Age	x_8	0.04

BMI (x_6)については，「糖尿病者」データにおいて各データ集合における要因間の変化が一番大きい．**図-3**と**図-4**で示しているように，「健常者」データにおいてのBMIは， Blood Pressure・Skin Thickness・Insulinと直接相関関係を持っている．しかし，「糖尿病者」データに対してはSkin Thicknessのみの関係を持ち，この結果からBMIは，今回使用した「健常者」データと「糖尿病者」データとの特徴を示す重要な要因の1つであると判断する．

次に，**表-2**の結果から分かるように，Ageに対する「健常者」データと「糖尿病者」データとの要因間の変化度が高い．これは，「健常者」データ $\mathrm{D_A}$ に対する相関グラフ（**図-3**）では，Glucoseとの相関を持っているが，「糖尿病者」データ $\mathrm{D_B}$ の相関グラフ（**図-4**）ではAgeとGlucose との相関性がなくなっていることから，データ $\mathrm{D_A}$ と $\mathrm{D_B}$ の構造に変化がみられる．したがって，Ageは「健常者」データと「糖尿病者」データとの特徴を示す重要な要因の一つであるとも考えらる．

Pregnancies (x_1) については，**表-3**からみると，BMIの次に「健常者」データ $\mathrm{D_A}$ に対する相関係数がと「糖尿病者」データにおける相関係数が大きく変わっている．しかし，**図-3**と**図-4**では，Ageだけの関係を持ち，変化はみられない．Pregnancies とAge との相関係数が0.44であり，Pregnanciesの相関変化にも影響を与えられるからである．これらの理由は，Ageとの相関が強いと判断される．

4. サポートベクターマシン

パターン認識の一手法であるサポートベクターマシン（Support Vector Machines, SVM）[5] は2つのクラスを明確に分けるために使用する．このとき，マージンを最大化することで汎化限界を最適化することができる．

すべての要因を用いる場合のSVMによる識別率は74%

である．そこで，要因を一つずつ抜き，残りの7要因を用いてSVMによる識別率（図-5）を求めることにより，各要因が識別率に与える寄与度について検証する．

BMIとGlucoseを取り除いた場合は，識別率に低くなり，糖尿病診断に対し，重要な要因であると考えられる．一方，Pregnancies と Diabetes Pedigree Functionに対しては，これらの要因がなくても識別率が75%であり，糖尿病診断に対し影響度は低いと判断できる．

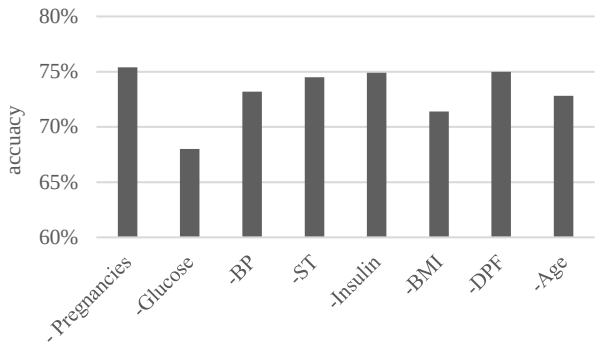


図-5　各要因の重要度（SVMによる識別率）

5. 結論

本研究では，グラフィカルラッソを用いて糖尿病データに対する健常者と糖尿病者の場合においてのデータ間の変化についての分析を行った．グラフィカルラッソによる疎な相関グラフ学習法を用いて，異なる2つのデータ集合間の場合に対する隣接行列から各要素（変数）間の相関変化度を導入することで，構造変化に与える主な要因による結果との比較検証より，本研究で得られた結果と確認した．しかし，グラフィカルラッソによる項目間の関係性と相関変化度は把握できたが，どの項目が診断において重要な要因になるかについては，分析方法によって異なるため，今後検討する必要がある．

参考文献

1) 日本人口動態統計　厚生労働省　2015
2) データヘルス　厚生労働省　2017
3) Pima Indians Diabetes Kaggle　2007 https://www.kaggle.com/
4) 疎な相関グラフの学習による相関異常の検出　井手剛　2009
5) Learning with Kernels - Support Vector Machines, Regularization, Optimization, and Beyond. Adaptive Computation and Machine Learning Series　B. Schoelkopf, A.J. Smola The MIT Press, Cambridge　2002

STUDY ON APPLICATION OF GRAPHICAL LASSO TO CHANGE DETECTION FOR DIABETES

Yeunkyung KIM, Yeboon YUN, Min YOON, Hirotaka NAKAYAMA

Major lifestyle-related diseases caused by unbalanced diet and irregular sleep patterns include diabetes, hypertension, and dementia. According to the demographic statistics of Japan for 2015, approximately 60% of total deaths are caused by the lifestyle-related diseases. In recent years, ‘Data Health’ has been increasingly conducted in Japan. With such projects, various data are analyzed through mathematical scientific methods and are used to enhance health, build preventive models, and detect diseases early for large populations.

In this study, using the graphical lasso, we try to analyze diagnosis data on diabetes, which is one of the most common diseases. Graphical lasso is a kind of sparse structure learning based on so-called precision matrices for two groups with different labels, and is applicable for detecting an anomaly among factors. The obtained results through the analysis will be expected to contribute for increasing efficiency in disease prevention and health promotion.